A MULTIBAND EXCITATION VOCODER AT 1.5 kb/s

by

NIMESH EASOW

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KANPUR

JANUARY 1997

A MULTIBAND EXCITATION VOCODER AT 1.5 kb/s

A Thesis Submitted

in Partial Fulfillment of the Requirements

for the Degree of

Master of Technology

by

Nimesh Easow

to the

Department of Electrical Engineering

Indian Institute of Technology, Kanpur

January, 1997.

30/1/97

CERTIFICATE

It is certified that the work contained in the thesis entitled **A Multiband Excitation Vocoder at 1.5kb/s** by **Nimesh Easow**, has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

Preeti Rao

Dr. Preeti Rao
Assistant Professor
Department of Electrical Engineering,
Indian Institute of Technology, Kanpur.

January, 1997

Acknowledgments

I would like to thank my guide Dr.Preeti Rao without whose help this thesis would not have been possible.I would also like to express my gratitude to Dr.Umesh for allowing me to use his PC for playing back the synthesised speech.

I take this oppurtunity to thank Prashant whose lab-keys I used most of the time and Apu for helping me out with LaTeX.

Thanks are also due to all my friends Davis,Saji,Prince,Prashant, Biju,Murali & George for making the life at IITK an enjoyable experience. Special thanks to Umesh ,the discussions with whom ,even late into the night, which I hope will not forget for a long time to come.I would also like to appreciate the co-operation showed by my class-mates Suyog,Arvind, Shaikh,Anand and Pratibha both during the course work and thesis.

Nimesh Easow

Abstract

Algorithms for the implementation of a Multiband Excitation vocoder at 2.4kb/s and 1.5 kb/s are presented. The MBE vocoder models the short time speech spectrum as the product of an excitation spectrum and a spectral envelope.Unlike in other vocoders ,where the excitation is specified by a fundamental frequency and a single v/uv decision, the excitation spectrum in MBE is represented by fundamental frequency and a v/uv decision for each harmonic of the fundamental frequency.The spectral envelope is represented by the samples at the fundamental frequency.During speech analysis the parameters of the spectrum are estimated in such a way that the synthetic spectrum is close to the original spectrum in the m. s. e sense.The excitation parameters are scalar quantised and the spectral envelope parameters are modelled using LPC spectrum and quantised in the LSF domain using an efficient split vector quantisation to obtain a speech coder operating at 2.4kb/s.We have used frame interpolation to further reduce the bit rate to 1.5kb/s.Adaptive post filtering is applied to the reconstructed speech.The coder has been simulated using C language.Informal listening reveals that the output speech is highly intelligible.While naturalness is well preserved at 2.4kb/s, there is a perceptible reverberance,particularly in female voices.The performance of the coder in the presence of additive noise is satisfactory. The 1.5kb/s coder provides a speech output that is slightly degraded in quality but otherwise comparable to that of the 2.4kb/s coder.

Contents

1 Introduction **1**

1.1 Review 1

1.2 The motivation for the refinement of excitation signal 2

1.3 Organisation of the thesis 3

2 Multi-Band Excitation Speech Model **4**

3 MBE Speech Analysis **6**

3.1 Introduction 6

3.2 Estimation of Pitch period and Spectral envelope parameters . 7

3.3 Look-Back Pitch Tracking 11

3.4 Look-Ahead Pitch Tracking 12

3.5 Estimation of v/uv information 15

3.6 Estimation of spectral amplitudes 16

4 Quantisation of MBE Parameters **18**

4.1 Introduction 18

4.2 Quantisation of fundamental frequency 19

4.3 Quantisation of v/uv decisions 19

4.4 Quantisation of Spectral Magnitudes 19

4.4.1 LPC modelling of MBE magnitudes 21
4.4.2 Quantisation of LPC model parameters 24
4.4.3 LSFs and their properties 24
4.4.4 Quantisation of LSF parameters 26
4.4.5 Split vector quantisation of LSF parameters 26
4.5 MBE at 1.5 kb/s 28

5 MBE Speech Synthesis 32
5.1 Introduction 32
5.2 Unvoiced speech synthesis 33
5.3 Voiced speech synthesis 35
5.3.1 Prediction of phases 37
5.4 Adaptive post filtering for speech enhancement 38

6 Coder Implementation and Performance 42

List of Figures

3.1 Initial Pitch Estimation window 9

3.2 Pitch Refinement window 10

3.3 Figure showing (a) Variation of error with pitch period (b) Original and Synthetic spectrum for correct pitch (c) Error and threshold for each band for v/uv decision 14

4.1 Block Diagram of MBE encoder at 2.4kb/s 20

4.2 Block Diagram of MBE decoder at 2.4kb/s 20

4.3 The LPC model generated using harmonic magnitudes 23

4.4 (a)Figure showing the spectrum and the LSFs (b)Figure showing the localised spectral sensitivity of LSFs 25

4.5 Interpolated Spectrum for Voiced speech 30

4.6 Interpolated Spectrum for Unvoiced speech 30

5.1 Synthesis Window 34

5.2 Block Diagram of Unvoiced Speech Synthesis 35

5.3 Block Diagram of Voiced speech synthesis 36

5.4 Figure showing reconstructed speech at various bit rates . . . 41

List of Tables

4.1 Bit allocation for INMARSAT coder 21
4.2 Bit allocation for 2.4kb/s coder 28
4.3 Bit allocation for 1.5kb/s coder 31

6.1 Processor time for 2.4kb/s coder 43
6.2 Processor time for 1.5kb/s coder 44
6.3 Algorithmic delay for 2.4kb/s coder 44
6.4 Algorithmic delay for 1.5kb/s coder 45

Chapter 1

Introduction

1.1 Review

Speech coding is the representation of speech signals in such a way that the number of bits for storage or transmission is reduced while maintaining the quality of the original signal. There are essentially two ways for representing speech signals.

1. Time domain representation or waveform representation;
2. Parametric representation which assumes an underlying method of speech production.

Waveform representation as the name implies is concerned with preserving the wave shape of the speech signal through a sampling and quantisation procedure.A wide range of coding methods like PCM,DM,ADPCM fall into this category.The typical bit rates of these coders are in the range of 128kb/s to about 20kb/s.

Further reduction in bit rate can only be achieved by exploiting the underlying model of speech production.These coders are based on the speech production model ,given by an excitation source driving a linear time varying filter.For this class of vocoders speech is analysed by first segmenting the speech using a time domain window. Then for each segment of the speech the excitation parameters and the system parameters are determined. The excitation parameters consist of the voiced/unvoiced decisions and the pitch period.The system parameters consist of the spectral envelope. In order to synthesise speech the excitation parameters are used to synthesise an excitation signal consisting of a periodic impulse train in the voiced regions and random noise in the unvoiced regions.The excitation signal is then filtered using the estimated system parameters.

1.2 The motivation for the refinement of excitation signal

Even though the vocoders consisting of this class of underlying speech models have been quite successful in synthesising intelligible speech,they have not been successful in synthesisng high quality speech.One of the major degradations present in vocoders employing a simple v/uv decision is a "buzzy" quality especially noticeable in regins of speech which contain mixed voicing or in voiced regions of noisy speech. Observations of the short time spectra indicate that these speech regions tend to have regions of the spectrum dominated by harmonics of the fundamental frequency and other regions dominated by noise like energy.Since speech synthesised entirely with a periodic source exhibits a "buzzy" quality and speech synthesised entirely with a noise source exhibits a "hoarse" quality ,it is clear that the perceived "buzziness" of the vocoder speech is due to replacing noise like energy in the original spectrum with periodic buzzy energy in the synthetic spectrum.This occurs since the simple v/uv excitation model produces excitation spectra consisting entirely of harmonics of the fundamental frequency or noise like energy.Since this problem is a major cause for quality degradation in vocoders, any attempt to significantly improve vocoder quality must account for these effects.

Inaccurate estimation of the model parameters has also been a major contributor to the poor quality of the synthesised apeech.

To summarise, the quality of the speech synthesised by the vocoder depends on two factors

1. complexity of the underlying model:

2. the accuracy of estimation of parameters of model,given the speech wave.

1.3 Organisation of the thesis

- Chapter 2 gives the MBE model,which offers an improvement in modelling the excitation signal.

- Chapter 3 gives methods for accurately estimating the parameters of this model.

- Chapter 4 gives methods for efficiently quantising these parameters.

- Chapter 5 gives methods for reconstructing the speeech signal given the MBE parameters.

- Chapter 6 gives the implementation details and the results obtained from implementation.

Chapter 2

Multi-Band Excitation Speech Model

Due to the quasi-stationary nature of the speech signal s(n), a window w(n) is usually applied to the speech signal to focus attention on a short time interval of approximately 20ms. The windowed speech signal is defined by

$$s_w(n) = w(n)s(n) \tag{2.1}$$

Over a short time interval, the Fourier transform $S_w(\omega)$ of the windowed speech signal can be modelled as the product of a spectral envelope $H_w(\omega)$ and an excitation spectrum $E_w(\omega)$,

$$\hat{S}_w(\omega) = H_w(\omega)E_w(\omega) \tag{2.2}$$

As in many simple speech models the spectral envelope $|H_w(\omega)|$ is a smoothed version of the original speech spectrum $|S_w(\omega)|$.The spectral envelope can be represented by linear prediction co-efficients,cepstral co-efficients, formant frequencies and bandwidths or samples of the original speech spectrum.The representational form of the spectral envelope is not the dominant issue in this model. However the spectral envelope must be represented accurately enough to prevent degradations in the spectral envelope from dominating the quality improvements achieved by the addition of a frequency dependent v/uv mixture function.

The excitation spectrum in MBE model differs from previous simple models in one major respect.In previous simple models the excitation spectrum is specified by the fundamental frequency ω_0 and a v/uv decision for the whole spectrum.In MBE model the excitation spectrum is specified by fundamental frequency ω_0 and a frequency dependent v/uv mixture function.In general, a continuously varying frequency dependent v/uv mixture function would require a very large number of parameters to represent it accurately.The addition of a large number of parameters would severely decrease the utility of this model in such applications as bit rate reduction.To reduce this problem ,the frequency dependent v/uv mixture function has been restricted to a frequency dependent binary v/uv decision.To further reduce the number of these binary parameters ,the spectrum is divided into a number of bands and a binary v/uv decision is allocated to each band.Due to the division of the spectrum into multiple frequency bands with a binary v/uv parameter for each band, this model is called the Multiband Excitation model.

The excitation spectrum $|E_w(\omega)|$ is obtained from the fundamental frequency ω_0 and the v/uv parameters by combining segments of a periodic spectrum $|P_w(\omega)|$ in the frequency bands declared voiced and a random noise spectrum in the frequency bands declared unvoiced.The periodic spectrum is completely specified by ω_0 .One method for generating the periodic spectrum is to take the Fourier transform magnitude of a windowed impulse train with pitch period P=$(2\pi/\omega_0)$. In another method the Fourier transform of the window is centered around each harmonic of the fundamental frequency and summed to produce the periodic spectrum.The v/uv decisions allows us to mix the periodic spectrum with a random noise spectrum in the frequency domain in a frequency dependent manner in representing the excitation spectrum.

Chapter 3

MBE Speech Analysis

3.1 Introduction

The basic question in any speech analysis based on a given model is : "Given the speech signal,how will we estimate the parameters of the model such that the error between the original speech signal and the synthetic one is minimised in some sense".

In many aproaches, the algorithms for the estimation of excitation parameters and estimation of spectral envelope parameters operate independently.These parameters are estimated based on some reasonable but heuristic criterion without explicit consideration of how close the synthesised speech will be to the original one[1].

In MBE analysis, the excitation and spectral envelope parameters are estimated simultaneously so that the synthesised spectrum is closest in the least square sense to the spectrum of the original speech.This approach can be viewed as an analysis by synthesis method. i. e. the parameters of the model are estimated by actually synthesising the speech (in the frequency domain) at the encoder.

Simultaneous estimation of all the speech model parameters is a computationally prohibitive problem.Consequently the estimation process has been divided into two major steps. In the first the pitch period and spectral envelope parameters are estimated to minimise the error between the orginal spectrum $S_w(\omega)$ and the synthetic spectrum $\hat{S}_w(\omega)$.Then the v/uv decisions are made based on the closeness of fit between the original and the synthetic spectrum at each harmonic of the estimated fundamental.

The parameters of MBE speech model can be estimated by minimising the following error criterion.

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} |S_w(\omega) - \hat{S}_w(\omega)|^2 d\omega \tag{3.1}$$

This error criterion was chosen instead of an error function with frequency dependent weighting because it leads to fairly simple expression for the samples A_m of the spectral envelope $H_w(\omega)$.

3.2 Estimation of Pitch period and Spectral envelope parameters

The objective is to choose the pitch period and spectral envelope parameters to minimise the error of 3.1. In general minimising this error over all parameters is a computationally expensive problem.However, for a given pitch period the spectral envelope parameters can be easily estimated.To show this we divide the spectrum into frequency bands centered at each harmonic of the fundamental frequency. We will model the spectral envelope as constant in this interval with a value of A_m.This allows the error (error criterion 3.1) in the interval around the m^{th} harmonic to be written as

$$E_m = \frac{1}{2\pi} \int_{a_m}^{b_m} \left[|S_w(\omega) - A_m E_w(\omega)|^2 \right] d\omega \tag{3.2}$$

where the interval $[a_m, b_m]$ is an interval with a width of the fundamental frequency centered on the m^{th} harmonic of the fundamental.The error E_m is minimised at

$$A_m = \frac{\int_{a_m}^{b_m} S_w(\omega)E_w^*(\omega)d(\omega)}{\int_{a_m}^{b_m} |E_w(\omega)|^2 d\omega} \tag{3.3}$$

For voiced frequency intervals,the envelope parameters are estimated by substituting the periodic transform $P_w(\omega)$ for the excitation transform in 3.3.Note that the A_m obtained has both magnitude and phase.An efficient method for obtaining a good approximation for the periodic transform $P_w(\omega)$ is to precompute samples of the Fourier transform of the window w(n) and centre it around the harmonic frequency associated with this interval.When creating the spectrum it is very important to adjust the position of $P_w(\omega)$ and size of the transform used to make sure that the peak of the window transform is centered on the harmonic and dies down to a very small value at $\pm 0.5\omega_0$ around each harmonic.

For unvoiced frequency intervals, the envelope parameters are estimated by substituting idealised white noise (unity across band) for $|E_w(\omega)|$ in 3.3.

Note that the estimation of A_m requires the knowledge of whether the m^{th} harmonic is voiced/unvoiced which is not yet known. To overcome this problem we will assume an excitation which is entirely voiced.The problem of estimation of the spectral envelope parameters can now be thought of as the problem of approximating the signal $s(t)$ in terms of L orthogonal signals $\{\phi_k(t)\}_{k=1}^{L}$ s. t. the error $E = |s(t) - \sum_{k=1}^{L} A_k\phi_k(t)|^2$ is minimised.That is, for each value of the pitch we will find the parameter A_k and calculate the error E.The parameters P and $\{A_k\}_{k=1}^{L}$ are chosen as those for which this error is minimised.

However there is a problem in choosing the error criteria above.Suppose that our speech segment has a spectra which contains small regions of pitch harmonics and large regions of noise like energy.The pitch period obtained above should reflect pitch harmonics in that small region.That is, the error

E should not vary with the pitch period for a spectrum consisting entirely of noise like energy.But since the L above is a function of the fundamental frequency ω_0, the error will be smaller for smaller ω_0 since L is larger.This bias for smaller ω_0 can be calculated [1] and an unbiased error criterion E_{UB} is developed by multiplying the error by a pitch period dependent correction factor to produce

$$E_{UB} = \frac{\int_{-\pi}^{\pi} |S_w(\omega) - \hat{S}_w(\omega)|^2}{(1 - P\sum_{-\infty}^{\infty} w^4(n)) \int_{-\pi}^{\pi} |S_w(\omega)^2 d(\omega)} d\omega \tag{3.4}$$

To obtain this result, the window w(n)(Figure 3.1) was normalised to have unit energy.The error E_{UB} has been normalised so that the minimum is near zero for a purely periodic signal and near one for a noise signal.This unbiased error criterion significantly improves the performance for noisy speech. In practice, these computations are performed by replacing integrals of continuous functions by summations of samples of these functions.

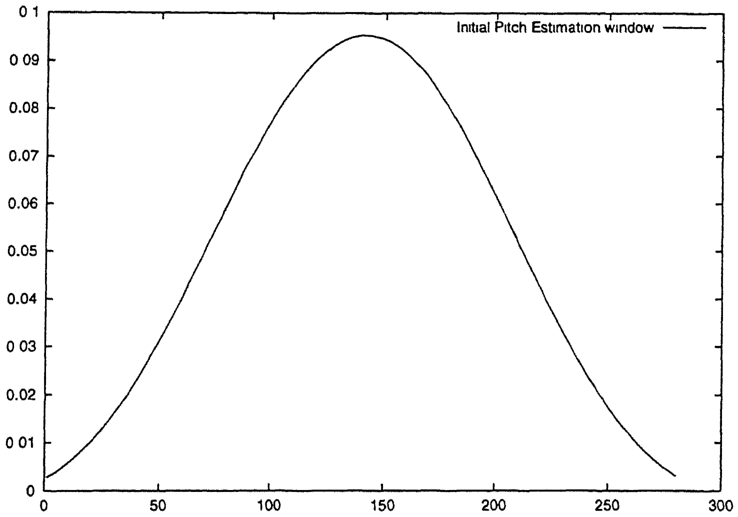

Figure 3.1: Initial Pitch Estimation Window

However evaluating the error criterion for all possible pitch periods can be computationally expensive. Reasonable approximations [1] lead to a sub-

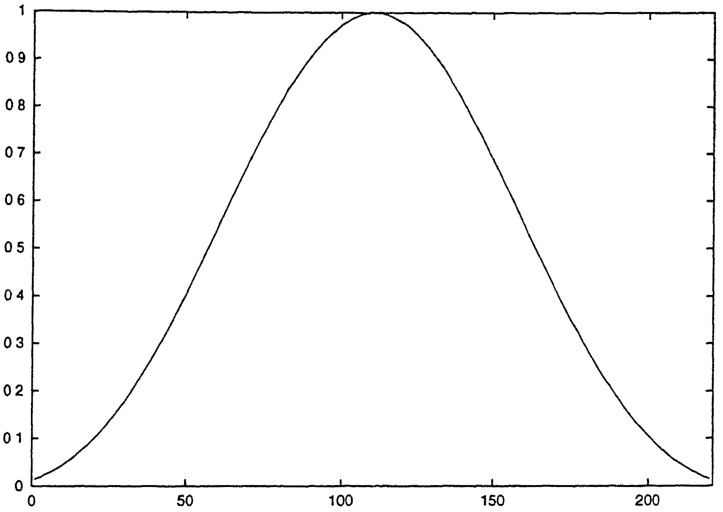

Figure 3.2: Pitch Refinement Window

stantially more efficient method for computing E_{UB}.

$$E_{UB} \approx \frac{\sum_{n=-\infty}^{\infty} w^2(n)s^2(n) - \psi(P)}{(1 - P\sum_{n=-\infty}^{\infty} w^4(n))(\int_{-\pi}^{\pi} |S_w(\omega)|^2 d\omega} \tag{3.5}$$

where

$$\psi(P) = P \sum_{k=-\infty}^{\infty} \phi(kP) \tag{3.6}$$

and $\phi(m)$ is the autocorrelation function of $w^2(n)s^2(n)$ given by

$$\phi(m) = \sum_{n=-\infty}^{\infty} w^2(n)s(n)w^2(n-m)s(n-m) \tag{3.7}$$

Minimising 3.5 over P is eqivalent to maximising 3.6.This technique is similar to the autocorrelation method but considers the peaks at multiples of the pitch period instead of only the peak at the pitch period.This suggest a computationally efficient method for obtaining an initial estimate of pitch period.

The initial pitch estimate P_I is chosen in such a way that $E(P_I)$ is small.How ever, P_I is not chosen simply to minimise $E(P)$. Instead pitch tracking must be used to account for pitch continuity between neighbouring speech frames.This is because if the pitch estimate is chosen such as to strictly minimise $E(P)$, then the pitch estimate may change abruptly between speech frames. This abrupt change in pitch can cause degradation in the synthesised speech.In addition pitch typically changes slowly. Therefore the pitch estimates from the neighbouring frames can aid in estimating the pitch of the current frame.

For each speech frame two different pitch estimates are computed. The first $\hat{P}_B$ is a backward estimate which maintains pitch continuity with previous speech frames. The second $\hat{P}_F$ is a forward estimate which maintains pitch continuity with future speech frames. These two estimates are compared with a set of decision rules[2] and either the forward estimate or the the backward estimate is chosen as the initial pitch estimate.

3.3 Look-Back Pitch Tracking

Let P_{-1} and P_{-2} denote the initial pitch estimates of the two previous speech frames.Let $E_{-1}(P_{-1})$ and $E_{-2}(P_{-2})$ denote the error functions of 3.5 obtained from the analysis of these two previous frames.

Since pitch continuity with previous frames is desired,the pitch for the current frame is considered in a range near P_{-1}.First,the error function $E(P)$ is evaluated at each value of P which satisfies the following constraints.

$$0.8P_{-1} \leq P \leq 1.2P_{-1} \tag{3.8}$$

$$P \epsilon \{21, 21.5, \ldots, 114\} \tag{3.9}$$

These values of E(P) are compared and P_B is defined as the value of P which satisfies these constraints and which minimises $E(P)$.The backward cumulative error is computed using the following formula.

$$CE_B(P_B) = E(P_B) + E_{-1}(P_{-1}) + E_{-2}(P_{-2}) \tag{3.10}$$

This is compared with the forward pitch estimate using a set of heuristics defined in the next section to determine the initial pitch estimate.

3.4 Look-Ahead Pitch Tracking

Look-ahead tracking attempts to preserve pitch continuity between future speech frames.Let $E_1(P)$ and $E_2(P)$ denote the error functions of equation 3.5 obtained from the two future speech frames.Since the pitch has not been determined for these future speech frames,the look ahead tracking algorithm must select the pitch for these future frames. This is done in the following manner.First P_0 is assumed to be fixed. Then the P_1 and P_2 are found which jointly minimise $E_1(P_1) + E_2(P_2)$,subject to the following constraints.

$$P_1 \epsilon \{21, 21.5, \ldots, 113.5, 114\} \tag{3.11}$$

$$0.8P_0 \leq P_1 \leq 1.2P_0 \tag{3.12}$$

$$P_2 \epsilon \{21, 21.5, \ldots, 113.5, 114\} \tag{3.13}$$

$$0.8P_1 \leq P_2 \leq 1.2P_1 \tag{3.14}$$

The values of P_1 and P_2 which jointly minimise $E_1(P_1) + E_2(P_2)$ subject to these constraints are denoted by $\hat{P}_1$ and $\hat{P}_2$ respectively.Once $\hat{P}_1$ and $\hat{P}_2$ have been computed the forward cumulative error function $CE_F(P_0)$ is computed according to:

$$CE_F(P_0) = E(P_0) + E_1(\hat{P}_1) + E_2(\hat{P}_2) \tag{3.15}$$

This process is repeated for each P_0 in the set $\{21, 21.5, \ldots 113.5, 114\}$. The corresponding values of $CE_F(P_0)$ are compared and $\hat{P}_0$ is defined as the value of P_0 in this set which results in the minimum value of $CE_F(P_0)$.

Once $\hat{P}_0$ has been found the integer submultiples of $\hat{P}_0$ must be considered.Every submultiple which is greater than or equal to 21 is computed and replaced with the closest member in the set $\{21, 21.5, \ldots, 113.5, 114\}$.

The smallest of these constraints is checked agaist constraints 3.16,3.17 and 3.18.If this submultiple satisfies any of these constraints then it is selected as the forward pitch estimate $\hat{P}_F$.Otherwise the next largest submultiple is checked against these constraints and it is selected as the forward pitch estimate if it satisfies any of these constraints.This process continues until all pitch submultiples have been tested against these constraints.If no pitch sub-multiple satisfies any of these constraints then $\hat{P}_F = \hat{P}_0$.

$$CE_F(\frac{\hat{P}_0}{n}) \leq 0.85 and \frac{CE_F(\frac{\hat{P}_0}{n})}{CE_F(\hat{P}_0)} \leq 1.7 \tag{3.16}$$

$$CE_F(\frac{\hat{P}_0}{n}) \leq 0.40 and \frac{CE_F(\frac{\hat{P}_0}{n})}{CE_F(\hat{P}_0)} \leq 3.5 \tag{3.17}$$

$$CE_F(\frac{\hat{P}_0}{n}) \leq 0.05 \tag{3.18}$$

Once the forward pitch estimate and the backward pitch estimate have both been computed the forward cumulative error and backward cumulative error are compared.Depending on the results of this comparison either $\hat{P}_F$ or $\hat{P}_B$ will be selected as the initial pitch estimate $\hat{P}_I$. The following decision rules is used to select the initial pitch estimate from these two candidates.

If

$$CE_B(\hat{P}_B) \leq 0.48, \text{then} \hat{P}_I = \hat{P}_B \tag{3.19}$$

Else if

$$CE_B(\hat{P}_B) \leq CE_F(\hat{P}_F), \text{then} \hat{P}_I = \hat{P}_B \tag{3.20}$$

Else

$$\hat{P}_I = \hat{P}_F \tag{3.21}$$

Since the autocorrelation domain method is some what less accurate than the frequency domain method described earlier , the frequency domain method is used to refine the initial coarse fundamental frequency estimate provided

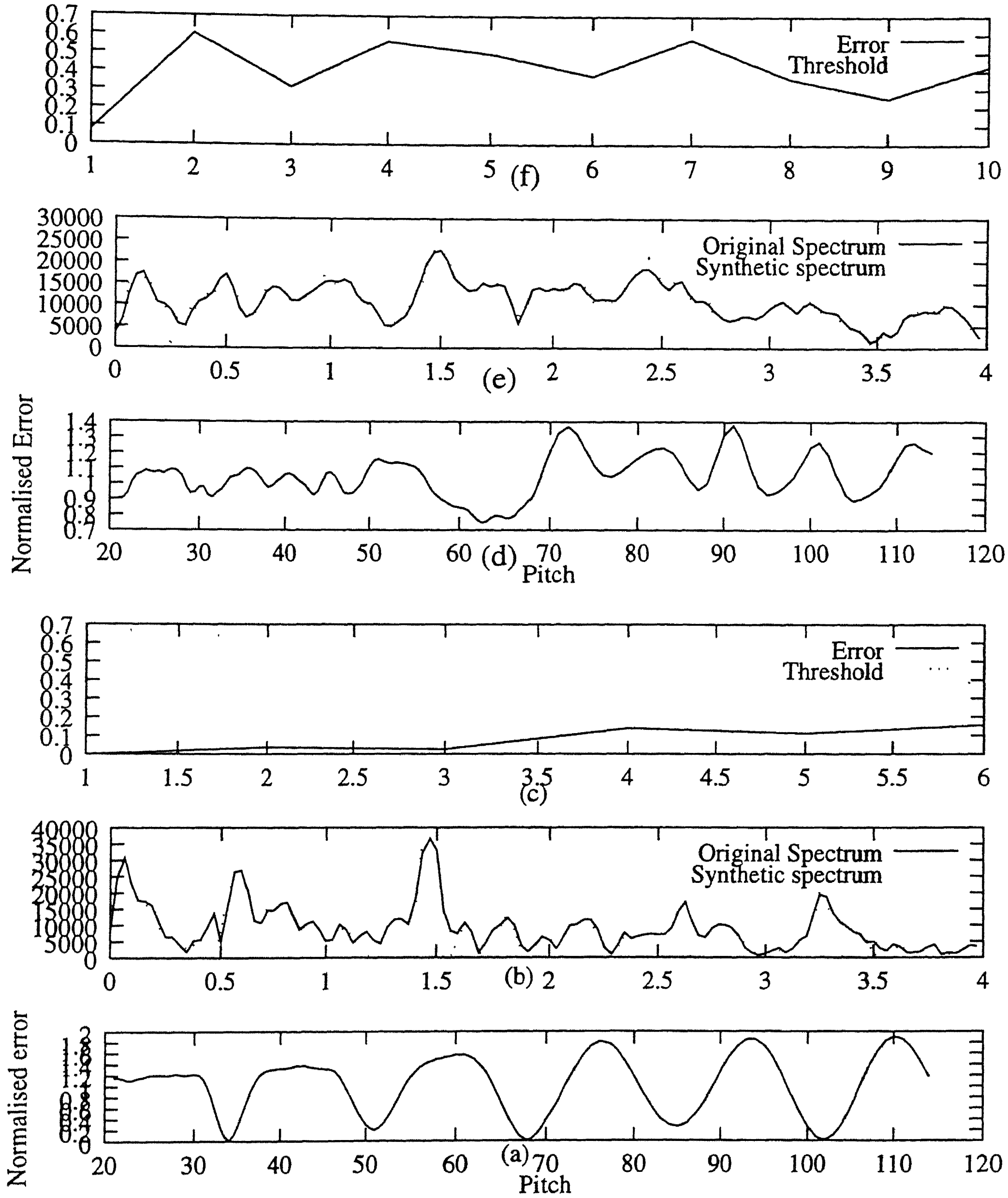

Figure 3.3: Figure showing for a voiced segment(a)Variation of error in eq (3.5) with pitch period (b)Original and synthetic spectrum for correct pitch (c)Error and threshold for each band for v/uv decisions(d),(e),(f) Figure showing the corresponding quantities for an unvoiced segment.

by the autocorrelation method.

3.5 Estimation of v/uv information

The v/uv determination is made by examining the normalised error D_k between the original and estimated speech spectra in each frequency band.

$$D_k = \frac{\sum_{n=b_{k_l}}^{b_{k_t}} |S_w(n) - \hat{S}(n, \omega_0|^2}{\sum_{n=b_{k_l}}^{b_{k_t}} |S_w(n)|^2} \tag{3.22}$$

where ω_0 is the refined fundamental frequency b_{k_l} and b_{k_t} are the first and last samples (in the frequency domain) of the k^{th} band, $S_w(n)$ is the original speech spectrum and $\hat{S}(n, \omega_0)$ is the reconstructed speech spectrum which is calculated using

$$\hat{S}(k, \omega_0) = A_l(\omega_0)W(k) \; 1 \leq l \leq L \; \lceil a_l \rceil \leq k \leq \lceil b_l \rceil \tag{3.23}$$

where $A_l(\omega_0)$ is the spectral amplitude calculated previously for the estimated fundamental frequency ω_0

$$\begin{aligned} W(k) &= \text{F.T of the window } w(n) \\ a_l &= (l - 0.5)\omega_0 \\ b_l &= (l + 0.5)\omega_0 \end{aligned}$$

The synthetic spectrum is assumed to be all voiced. However the speech spectrum is not all voiced and thus although the synthetic spectrum will be very similar to the original spectrum in the voiced regions it will have large differences in the unvoiced regions.This similarity or dissimilarity measure can be used to make the v/uv decision by comparing it against a predetermined threshold.

To determine the v/uv decisions, the normalised error,D_k, for each frequency band is compared with the adaptive threshold $\triangle_k(\omega_0)$ given by [2]

$$\triangle_k(\omega_0) = (\alpha + \beta\omega_0)[1 - \epsilon(k-1)\omega_0]M(\zeta_0, \zeta_{avg}, \zeta_{min}, \zeta_{max}) \tag{3.24}$$

where $\alpha = 0.35, \beta = 0.557$ and $\epsilon = 0.4775$ are factors that give good subjective quality and

$$M(\zeta_0, \zeta_{avg}, \zeta_{min}, \zeta_{max}) = \begin{cases} 0.5 & \zeta_{avg} < 200 \\ \frac{(\zeta_0+\zeta_{min})(2\zeta_0+\zeta_{max})}{(\zeta_0+\mu\zeta_{max})(\zeta_0+\zeta_{max})} & \zeta_{avg} \geq 200 \text{ and } \zeta_{min} < \mu\zeta_{max} \\ 1.0 & \text{otherwise} \end{cases}$$

μ is the adaptation factor that controls the decision threshold for v/uv decisions. The parameter ζ_0 is the avarage energy of the current speech frame.

The other parameters are updated every speech frame as follows [2]

$$\zeta_{avg}(j) = 0.7\zeta_{avg}(j-1) + 0.3\zeta_0 \tag{3.25}$$

$$\zeta_{max}(j) = \begin{cases} 0.5\zeta_{max}(j-1) + 0.5\zeta_0 & \text{if } \zeta_0 > \zeta_{max}(j-1) \\ 0.99\zeta_{max}(j-1) + 0.01\zeta_0 & \text{otherwise} \end{cases} \tag{3.26}$$

$$\zeta_{min}(j) = \begin{cases} 0.5\zeta_{min}(j-1) + 0.5\zeta_0 & \text{if } \zeta_0 \leq \zeta_{min}(j-1) \\ 0.975\zeta_{min}(j-1) + 0.025\zeta_0 & \text{if } \zeta_{min}(j-1) \leq \zeta_0 \leq \zeta_{min}(j \\ 1.025\zeta_{min}(j-1) & \text{otherwise} \end{cases} \tag{3.27}$$

A voicing decision for each band is made by comparing the normalised error for each band with the value of the threshold function. If the normalised error is less than the threshold function, the corresponding frequency band is declared voiced, otherwise unvoiced.

3.6 Estimation of spectral amplitudes

The spectral amplitudes are calculated differently for voiced and unvoiced parts of the spectrum.Since the synthetic spectrum was created by assuming voiced speech throughout the excitation spectrum, $A_l(\omega_0)$ calculated previously is assumed to characterise the spectral envelope for voiced speech accurately.

That is ,$M_l = |A_l(\omega_0)|$

The unvoiced harmonic magnitudes are represented by the r. m. s. value of speech in each unvoiced harmonic frequency region calculated using

$$M_l = \frac{1}{\sum_n w(n)} \sqrt{\frac{1}{\omega_0} \sum_{\lceil a_l \rceil}^{\lceil b_l \rceil} |S_w(n)|^2} \tag{3.28}$$

To summarise the analysis algorithm is presented

1. Window a speech segment using the analysis window (Typically 20ms duration Hamming window).
2. Compute the unbiased error criterion versus pitch period using the efficient autocorrelation domain approach.The error is computed for all values of pitch in the $\{21, 21.5, \ldots, 113.5, 114\}$.
3. Use pitch tracking to maintain continuity with previuos frames.
4. Refine this pitch by using the more accurate frequency domain method.
5. Estimate the spectral parameters A_l by assuming the excitation spectrum to be all voiced.
6. Make a v/uv decision for each frequency band in the spectrum.
7. The final spectral envelope parameter representation is composed by combining voiced spectral envelope parameters in those frequency bands declared voiced with unvoiced spectral envelope parameters in those frequency bands declared unvoiced.

Chapter 4

Quantisation of MBE Parameters

4.1 Introduction

The speech synthesis at the MBE decoder requires information about the fundamental frequency,v/uv decisions,spectral magnitudes and phases of the voiced harmonics. Since the phase information of the voiced harmonics can be predicted,no phase information is sent between encoder and decoder.

Although pitch is estimated at quarter sample accuracy in the time domain (this being done so as to estimate spectral ampli tudes accurately) it is quantised with half sample accuracy using 8 bits . The v/uv information is binary and requires no quantisation. The set of spectral amplitudes requires much attention for accurate and bit efficient quantisation.Each of these is discussed in detail below.

4.2 Quantisation of fundamental frequency

The values of the pitch is restricted to the set $\{21, 21.5 \ldots, 114\}$. This parameter is uniformly quantised using 8 bits enabling half sample accuracy.The following equation may be used to quantise pitch [2]

$$b_0 = \lfloor \frac{4\pi}{\omega_0} - 39 \rfloor \tag{4.1}$$

and decoded using

$$\omega_0 = \frac{4\pi}{b_0 + 39.5} \tag{4.2}$$

4.3 Quantisation of v/uv decisions

The v/uv decisions are binary,therefore they can be encoded using one bit per decision band, the maximum number of decision bamds being restricted to 12.The harmonics beyond the coverage of 12 v/uv bands are treated as unvoiced.The v/uv decisions can be encoded using [2]

$$b_1 = \sum_{k=1}^{K} v_k 2^{K-k} \tag{4.3}$$

and decoded using

$$v_k = \lfloor \frac{b_1}{2^{K-k}} \rfloor - 2\lfloor \frac{b_1}{2^{K+1-k}} \rfloor, 1 \leq k \leq K \tag{4.4}$$

where K is the number of bands in the current frame.

4.4 Quantisation of Spectral Magnitudes

In a typical MBE coder, most of the bits are allocated to the quantisation of the spectral magnitudes.In the case of INMARSAT-M [2] system where the source coder operates at 4.15kb/s with a 50Hz frame rate , only 20(8+12) bits

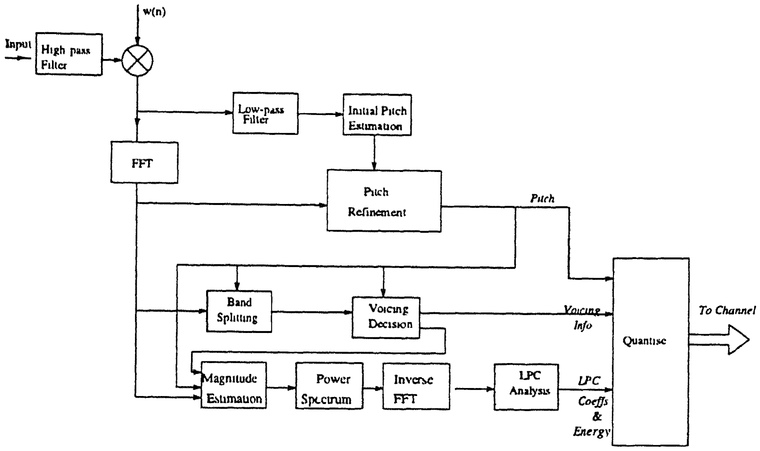

Figure 4.1: Block Diagram of MBE encoder at 2.4 kb/s.

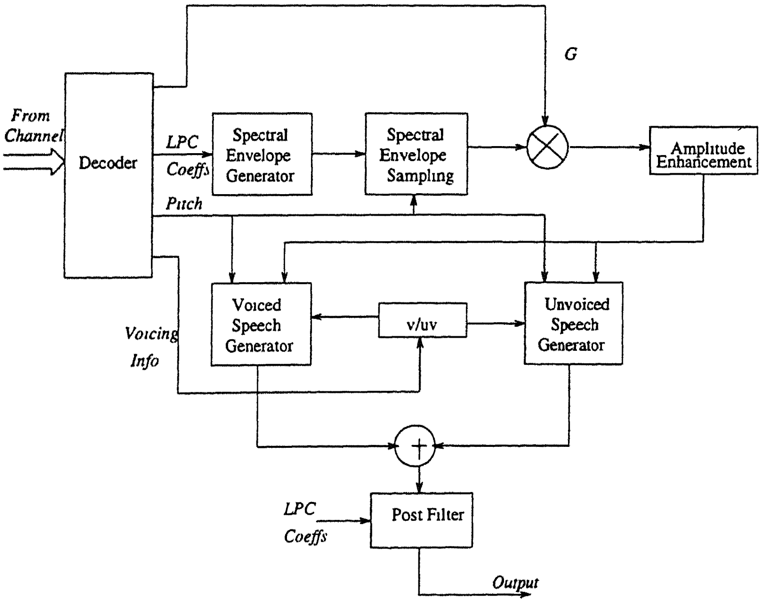

Figure 4.2: Block Diagram of MBE decoder at 2.4kb/s.

were used for the quantisation of pitch and v/uv information and the rest of the bits(63) were used to quantise the spectral magnitudes.Since an accurate pitch is necessary for speech synthesis and 12 v/uv bands are just enough to cover the 4kHz speech spectrum, the overall bit rate of the coder can be significantly reduced only by employing more efficient magnitude quantisation schemes.

Table 4.1: **Bit allocation for INMARSAT coder**

Parameter	*No: of bits*
Pitch	8
v/uv	12
Spectral amplitudes	63

Reduction in bit rate can be achieved by employing vector quantisation of all spectral amplitudes.However since the number of spectral amplitudes per frame is a function of fundamental frequency (ω_0), it is very difficult to design a code book to match this variation. Therefore it is more practical to vector quantise the general shape of the spectrum ,which is independent of fundamental frequency.

4.4.1 LPC modelling of MBE magnitudes

Since the spectral magnitudes can be thought of as the samples of the spectral envelope and since what we want to represent is the general shape of the spectrum, it is very sensible to model the spectral magnitudes using a 10th order all pole LPC filter. Such an all pole filter can be written as

$$H(\omega) = \frac{G}{A(\omega)} = \frac{G}{1 + \sum_{k=1}^{p} a_k e^{-jk\omega}} \tag{4.5}$$

where G =gain p =filter order. The m.s.e E_r between $S(\omega)$ and $H(\omega)$ is therefore given by

$$E_r = \sum_{n=0}^{N+p-1} e_r^2(n) = \sum_\omega |S_\omega(\omega)|^2 |A(\omega)|^2$$

$$= G^2 \sum_\omega \left(\frac{|S_\omega(\omega)|}{|H(\omega)|} \right)^2 \tag{4.6}$$

The parameters $\{a_k\}$ are determined by minimising E_r w. r. t. a_k, i. e. $\{\frac{\delta E_r}{\delta a_k} = 0\}_{k=1}^{p}$. These conditons reduce to

$$\sum_{k=1}^{p} a_k R_{|\imath-k|} = -R_\imath \tag{4.7}$$

where $R_k = \sum_\omega |S_\omega(\omega)|^2 \cos(k\omega)$

Equation 4.7 gives a set of p linear equations in p unknowns which can be solved for $\{a_k\}$

Similarly G can be calculated using

$$G^2 = R_0 + \sum_{k=1}^{p} a_k R_k \tag{4.8}$$

In the above analysis, the original speech spectrum $S(\omega)$ was matched with the defined LPC model. However in the MBE model only the magnitudes at the harmonics of the fundamental frequency are available for the spectral envelope estimation. The error E_r and the auto-correlation coefficients should therefore be redefined to consider the envelope at harmonically spaced frequencies [3].

$$E_r = \frac{G^2}{L} \sum_{k=0}^{L} \frac{|S_\omega(k\omega_0)|^2}{|H(k\omega_0)|^2} \tag{4.9}$$

and

$$R_k = \frac{1}{L} \sum_{l=0}^{L} |S_\omega(l\omega_0)|^2 \cos(kl\omega_0) \tag{4.10}$$

L=number of spectral points.

Since the number of harmonics in the spectrum is a function of the funda-

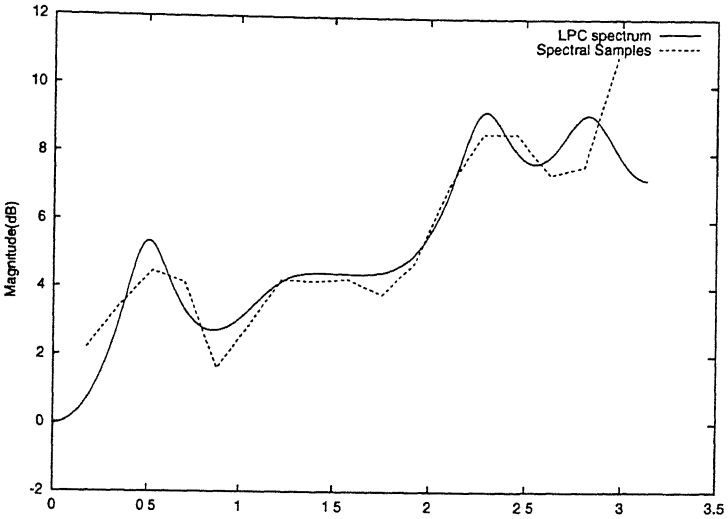

Figure 4.3: The LPC model generated using harmonic magnitudes.

mental frequency,sparse sampling the spectral envelope before LPC modelling can have a varying accuracy.This can be a significant problem for high pitch voices.To increase the number of spectral points linear interpolation in the log-domain can be done.The interpolation was done such that 8 more points were added between two spectral amplitudes. i. e. the interpolated spectrum was calculated at $k\omega_0 + l\omega_0/8$ $1 \leq k \leq L, 1 \leq l \leq 8$

The interpolated spectrum $Q(\omega)$ is computed using

$$\log_{10}|Q(\omega)| = \log_{10}|S_\omega(\omega_k)| + \left(\frac{\omega - \omega_k}{\omega_{k+1} - \omega_k}\right)[\log_{10}|S_\omega(\omega_{k+1}) - \log_{10}|S_\omega(\omega_k)|] \tag{4.11}$$

for $\omega_k \leq \omega \leq \omega_{k+1}$ where ω_k is the frequency of the k^{th} harmonic.

Given an original spectrum $S(\omega)$ and model spectrum $H(\omega)$ known at L frequency points,the accuracy of spectral fit between them can be measured using the spectral distortion given by

$$SD = \frac{1}{L}\sum_{k=1}^{L}[10\log_{10}|S(k\omega_0)| - 10\log_{10}|H(k\omega_0|]^2 \tag{4.12}$$

An average spectral distortion of 2.8dB was obtained using the above distortion measure.

4.4.2 Quantisation of LPC model parameters

For purposes of quantisation two desirable properties for a parameter set to have are

1. filter stability upon quantisation
2. a natural ordering of the parameters.

Property 1 means that the poles of H(z) continue to be inside the unit circle even after parameter quantisation. By 2 we mean that the parameters exhibit an inherent ordering. For e. g. in the parameter set $a_1, a_2, \ldots, a_p$ H(z) is not the same in general if a_1 and a_2 are interchanged.When an ordering is present, a statistical study on the distribution of individual parameters can be made to develop better encoding schemes.

It is clear that the LPC parameters does not satisfy property 1 and therefore is not suitable for quantisation.For this we will have to consider a transformation of the LPC parameters so that it will have these two properties.Traditionally it has been known that the reflection coefficients k_i exhibit these two properties. In the next section we study another representation of LPC parameters i. e. line spectral frequencies (LSF) and present a property (localised spectral sensitvity) not possesed by reflection coefficients which make them ideal for split vector quantisation.

4.4.3 LSFs and their properties

First we will define LSFs. The inverse filter A(z) is used to construct two polynomials

$$P(z) = A(z) + z^{-(M+1)}A(z^{-1}) \tag{4.13}$$

$$Q(z) = A(z) - z^{-(M+1)}A(z^{-1}) \tag{4.14}$$

The roots of the polynomials P(z) and Q(z) are called the LSFs. The polynomials P(z) and Q(z) have the following properties

1. All zeroes of P(z) and Q(z) lie on the unit circle.

2. Zeroes of P(z) and Q(z) are interlaced with each other i. e the LSFs

are in ascending order. It can be shown that A(z) has the minimum phase property if it satisfies these two properties[4, 5].Thus the stability of the LPC synthesis filter can easily be ensured by quantising the LPC information in the LSF domain. A cluster of (2 to 3) LSFs characterises a formant frequency and the bandwidth of a given formant depends on the closeness of the corresponding LSFs. In addition the spectral sensitivities of the LSFs are localised, i.e a change in given LSF produces a change in the LPC power spectrum only in its neighbourhood.

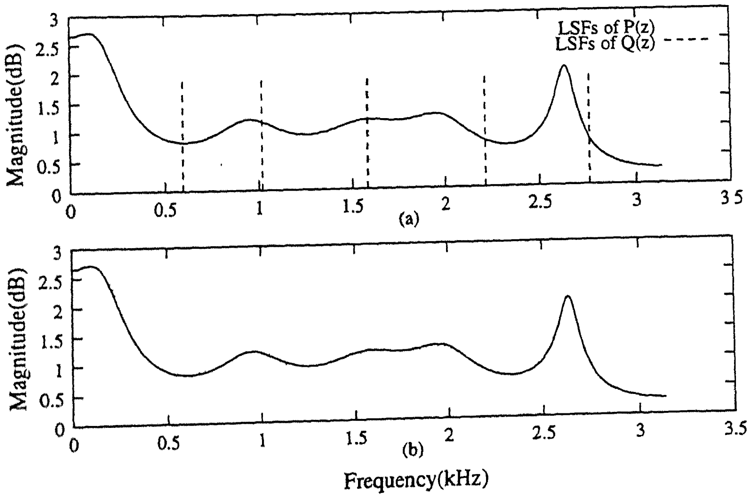

Figure 4.4: (a)Figure showing the spectrum and the LSFs (b)Figure showing the localised spectral sensitvity of LSFs when the LSF at 1.01 was changed to 1.035

The localised spectral sensitivity property of LSFs makes them ideal for split vector quantisation since the individual parts of an LSF vector can be independently quantised without leakage of quantisation distortion from one spectral region to another.This property also helps in giving different weights to different LSFs in a LSF based distance measure which might be useful as some LSFs are more important than others.

4.4.4 Quantisation of LSF parameters

It is well known that a vector quantiser results in smaller quantisation distortion than the scalar quantiser at any given bit rate. But if we vector quantise all the LSF parameters simultaneously, it will result in a code book having a very large number of code vectors.Such a vector quantiser will have the following problems [6]

1. A large code book requires prohibitively large amount of training data and the training process can take too much of computation time.
2. The storage and computational requirement in vector quantisation encoding will be prohibitively high.

Because of these problems , a suboptimal vector quantizer has to be used. In this work we have used split vector quantiser as the suboptimal quantiser.

4.4.5 Split vector quantisation of LSF parameters

In split vector quantisation ,the LPC parameter vector (in some suitable representation as LSFs or reflection coefficients) is split into a number of parts and each part is quantised seperately using vector quantisation. In order to design an optimal split vector quantiser the following questions must be answered.

1. For vector splitting what LPC representation should be used?

2. What distortion measure should be used?
3. In how many parts should the vector be split?
4. How many elements should there be in each of these parts?
5. How many bits should be allocated to each part?

We will consider these questions one by one. Because of the properties mentioned above the LSF representation has been found to be the best for split vector quantisation.The LSF vector was split into two parts,the first part containing four elements and the second part containing six elements.Each part is allocated 12 bits each.The choice of the distortion criterion is the most important criteria in the design of vector quantiser and is discussed in the next section.

Distortion Measure

A weighted Euclidean distance measure $d(\mathbf{f}, \hat{\mathbf{f}})$ between the test LSF vector $\mathbf{f}$ and the reference LSF vector $\hat{\mathbf{f}}$ is given by [4]

$$d(\mathbf{f}, \hat{\mathbf{f}}) = \sum_{i=1}^{10}[w_i(f_i - \hat{f}_i)^2] \tag{4.15}$$

where f and f_i are the i^{th} LSFs in the test and refernce vector respectively and w_i is the weight assigned to the i_{th} LSF.It is given by

$w_i = [P(f_i)]^r$

where $P(f)$ is the LPC power spectrum asscociated with the test vector as a function of frequency f and r is an empirical constant(taken as 0.15) which controls the relative weights given to different LSFs.

In the weighted Euclidean distance measure the weight assigned to a given LSF is propotional to the value of the LPC power spectrum at this LSF.Thus the distance meassure allows for quantisation of LSFs in the formant region better than those in the non-formant regions. How this distance measure is well suited for speech enhancement is seen in a later section.

Note that the choice of the distortion measure realises an efficient quantiser but not an optimal one in the sense of reducing the spectral distortion.The vector quantiser maps each vector $\mathbf{x}$ into $\hat{\mathbf{x}}$ s. t. the distortion measure $d(\mathbf{x}, \hat{\mathbf{x}})$ is minimised.But our distortion measure reduces the distance between the two LSFs without consideration of how the spectrum changes for a given change in the LSFs. That is ,we did not incorporate the spectral sensitivities of the LSFs into the distortion measure.Thus this will be optimal only if the spectral sensitivity is uniform.The idea is that there may exist a transformation (similar to the log-area ratio for reflection coefficients) that will make this quantiser optimal.

Each of the split vector quantiser was implemented using the LBG algorithm [7]. This completes the quantisation of LPC parameters.

Table 4.2: **Bit allocation for 2.4kb/s coder**

Parameters	*Bits/Frame*	*Bit Rate(kb/s)*
Pitch	8	0.4
10 LSFs	24	1.2
Energy	5	0.25
v/uv	12	0.60
Total	49	2.425

Bit allocation for 2.4 kb/s coder is shown in table 4.2.

4.5 MBE at 1.5 kb/s

Further reduction in the bit rate of a MBE coder can be achieved either by using a reduced frame rate or by using frame interpolation.In this work we have used frame interpolation to further reduce bit rate. Of all the MBE parameters the most important is the fundamental frequency.Therefore, to maintain good speech quality the fundamental frequency must be transmitted for each frame.The other parameters (LPC filter coefficients,the energy and v/uv decisions) can be transmitted in alternate frames. At the decoder ,the untransmitted parameters are estimated from the neighbouring transmitted sets by means of interpolation. Transmitted parameters are quantised and coded as in the 2.4kb/s version above. To achieve the best performance at the decoder ,the interpolation performance is measured at the encoder and some indication of how the interpolation should be performed is sent to the decoder. In the case of spectral envelopes an attempt is made to minimise the error

$$E = \sum_{\omega} |H_m(\omega) - \hat{H}_m(\omega)|^2 \tag{4.16}$$

where $H_m(\omega)$ is the original estimated spectral envelope of the m^{th} frame that is not transmitted and $\hat{H}_m(\omega)$ is the interpolated envelope using the two neighbouring frame envelopes.The envelope interpolation can be easily done in the LSF domain with appropriate contributions from each neighbouring LSF set. The current interpolated LSF coefficients can therefore be estimated as [3]

$$lsf_m(i) = lsf_{m-1}(i) + [lsf_{m+1}(i) - lsf_{m-1}(i)]\frac{k}{M-1}, k = 0, 1, 2 \ldots M-1 \tag{4.17}$$

where M is an integer that is a power of 2 (taken as 16). The energy of the current frame is estimated using a similar procedure. The current frame energy is reconstructed using

$$G_m = G_{m-1} + (G_{m+1} - G_{m-1})\frac{k}{M-1} \tag{4.18}$$

where G_m and G_{m+1} are the energies for the current and previous frames respectively. The spectral envelopes with their reconstructed energy levels are compared with original spectral envelope with its actual energy level. The index for the best spectral envelope $k_{best} = k$ which minimises error E_k is coded and transmitted to the receiver.

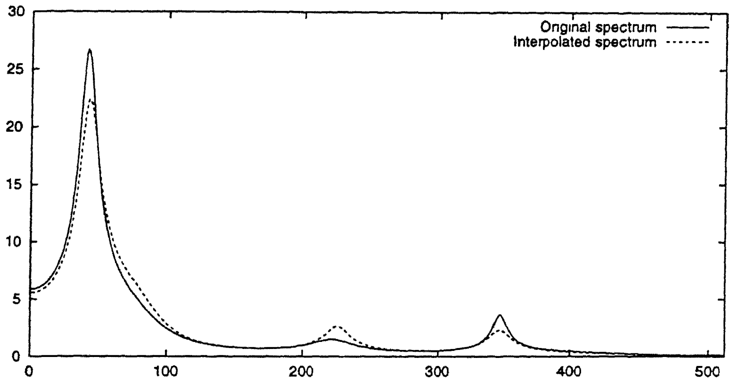

Figure 4.5: Interpolated Spectrum for Voiced speech.

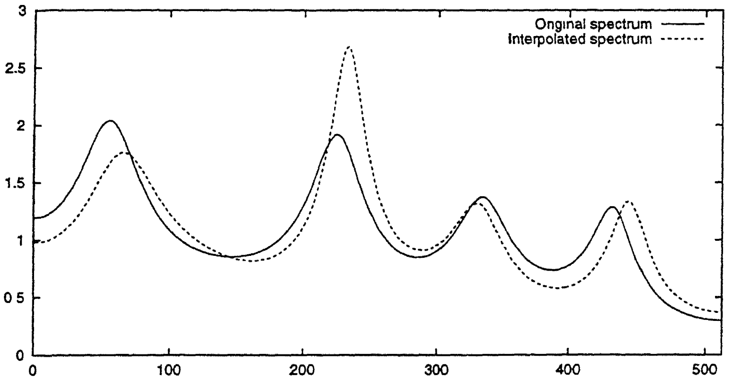

Figure 4.6: Interpolated Spectrum for Unvoiced speech.

The bit allocation for 1.5 kb/s coder is shown in table 4.3.

The v/uv information is estimated using a different method from the spectral envelope estimation. The v/uv estimation method can be formulated as [3]

$$v_m(i) = \begin{cases} v_{m-1}(i) & \text{if } k_{best} < M/2 \\ v_{m+1}(i) & \text{if } k_{best} \geq M/2 \end{cases}$$

where $v_m(i)$ and $v_{m+1}(i)$ are the i^{th} band v/uv decisions of the previous and the next frames. This is because $k_{best} > M/2$ means that the current frame is more close to the next frame than to the previous frame and vice

Table 4.3: Bit allocation for 1.5kb/s coder

Parameters	$(m-1)^{th}$ *Frame*	m^{th} *Frame*	*Bit rate(kb/s)*
Pitch	8	8	0.45
10 LSFs	24	-	0.6
Energy	5	-	0.125
v/uv	12	-	0.3
Index,k_{best}	-	4	0.10
Total	49	12	1.575

versa.

Chapter 5

MBE Speech Synthesis

5.1 Introduction

In the previous sections, the MBE model parameters and methods to estimate and quantise them were discussed.In this section ,an approach to synthesising speech from the model parameters is discussed.

There are essentially two methods for synthesising speech given the model parameters. One is the frequency domain approach and the other is the time domain approach.

In the frequency domain approach ,an excitation transform is constructed by combining segments of a periodic transform in frequency bands declared voiced with segments of a noise transform in frequency bands declared unvoiced.A spectral envelope is constructed by linearly interpolating between samples $|A_m|$. This seems to be the natural choice given the MBE model.But a problem can occur with this method when voiced speech is synthesised.Since the portion of the synthesised speech is modeled as a periodic signal with a constant fundamental over the entire frame, a large change in the fundamental frequency from one frame to the next can cause time discontinuities at the frame edges causing significant degradations of synthetic speech quality.

In the time domain approach, the voiced and unvoiced portions are synthesised seperately in the time domain and then added together.The voiced signal can be synthesised as the sum of sinusoidal oscillators with frequencies at the harmonic of the fundamental and amplitudes set by the spectral envelope parameters.This technique has the advantage of allowing a continuous variation in fundamental frequency from frame to frame,eliminating the problem of time discontinuities in the harmonics. The unvoiced part of the speech can be synthesised as the sum of outputs of bandpass filtered white noise.

The time domain method was selected for synthesising the voiced portion of the synthetic speech.This method was selected due to its advantage of allowing a continuous variation in fundamental frequency from one frame to the next.The frequency domain method was selected for synthesising unvoiced speech.This method was selected due to the ease of implementing a filter bank in the frequency domain using FFT algorithm.

5.2 Unvoiced speech synthesis

For each speech frame, a block of random noise u(n) is windowed and transformed with a fast fourier transform (FFT) to get $U(\omega)$. The noise sequence used for the implementation is given by

$$u(n+1) = 171u(n) + 11213 - 53125\left\lfloor\frac{171u(n) + 11213}{53125}\right\rfloor \tag{5.1}$$

The noise sequence is initialised to u(-105)=3147. The regions of the spectrum which corresond to voiced harmonics are set to zero.The remaining spectral components which correspond to the unvoiced part of the speech are then normalised to the unvoiced harmonic magnitudes.

$$\hat{U}(m) = \frac{M_l \gamma_w U(m)}{\sqrt{\sum_{n=\lceil a_l \rceil}^{\lceil b_l \rceil - 1} |U(n)|^2 / (\lceil b_l \rceil - \lceil a_l \rceil)}} \tag{5.2}$$

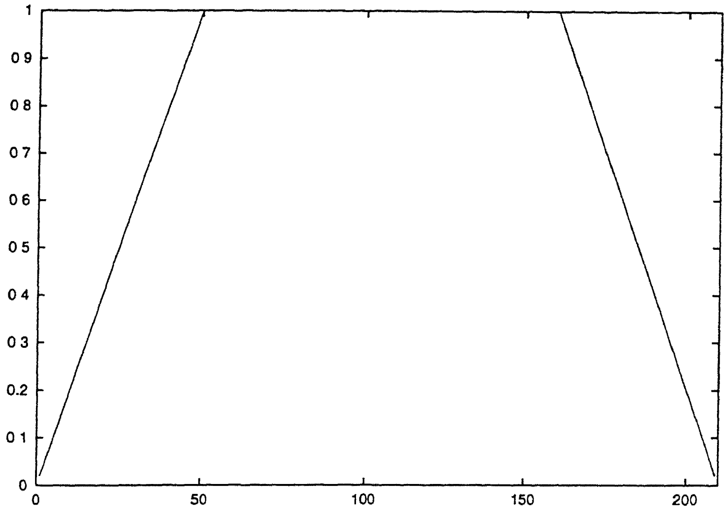

Figure 5.1: Synthesis Window

M_l = spectral amplitude of the l^{th} harmonic
where γ_w is a fixed weighting factor given by

$$\gamma_w = \left[\sum_n w_R(n)\right] \sqrt{\frac{\sum_n w_S^2(n)}{\sum_n w_R^2(n)}} \tag{5.3}$$

where $w_S(n)$ is the synthesis window and $w_R(n)$ is the pitch refinement(analysis) window.The phase in these regions is not modified and therfore corresponds to the phase of the original noise sequence.The inverse transform of this modified noise spectrum $\hat{U}(\omega)$ corresonds to the unvoiced part of the speech $\hat{u}(n)$, for that frame.However, since the length of the synthesis window is longer than the frame size,the unvoiced speech for each segment overlaps the neighbouring frames. The weighted overlap add procedure is used to average these sequences in the overlapping regions [8]

$$s_{uv}(n) = \frac{w_S(n)\hat{u}(n, j-1) + w_S(n-N)\hat{u}(n, j)}{w_S^2(n) + w_S^2(n-N)} 0 \leq n < N \tag{5.4}$$

where $\hat{u}(n, j)$ corresponds to the n^{th} sample in the j^{th} frame.

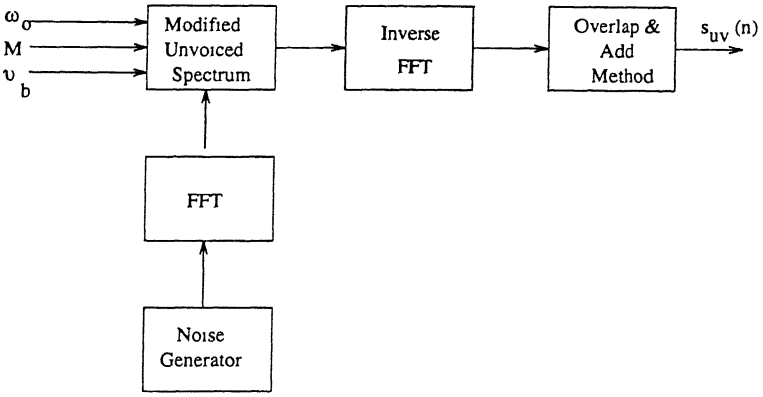

Figure 5.2: Block Diagram of Unvoiced Speech Synthesis.

5.3 Voiced speech synthesis

As already indicated synthesiser for voiced speech is implemented as a bank of tuned oscillators.However due to the fact that voiced portion of the speech is not perodic over intervals consisting of several analysis frames,the variations in the estimated parameters at adjacent frames can cause discontinuities at the edges of the frame ,resulting in significant degradation in speech quality.To overcome this problem during synthesis ,both current and previous frame parameters are checked to make sure a smooth transition at the frame boundaries take place.

Each harmonic oscillator is therefore implemented using the following rules[2].In the following equations ,$v_l(j)$ is the l^{th} harmonic in the j^{th} frame.0 represents unvoiced harmonic and 1 represents voiced harmonic.

If $v_l(j) = 0$ and $v_l(j-1) = 1$

$$s_v(n) = \sum_{l=1}^{L} w_S(n) M_l(j-1) \cos[\omega_0(j-1)nl + \phi_l(j-1)] \tag{5.5}$$

If $v_l(j) = 1$ and $v_l(j-1) = 0$

$$s_v(n) = \sum_{l=1}^{L} w_S(n-N) M_l(j) \cos[\omega_0(j)(n-N)l + \phi_l(j)] \tag{5.6}$$

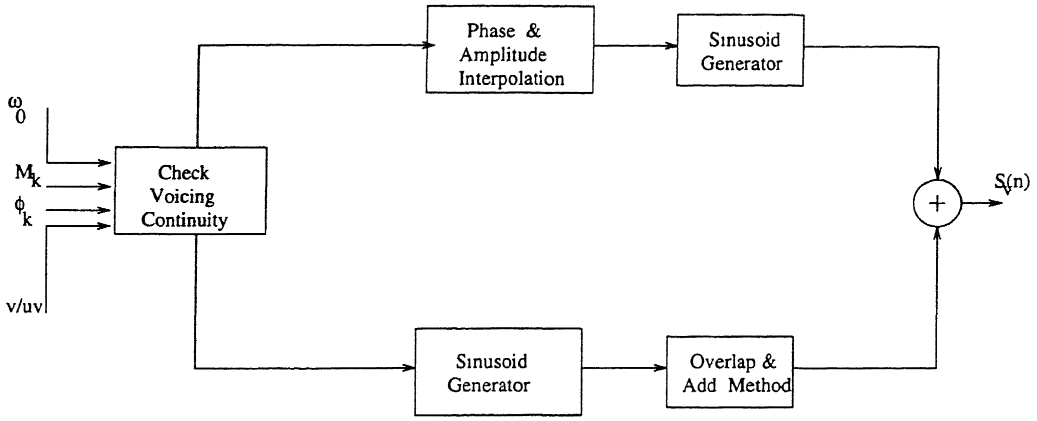

Figure 5.3: Block Diagram of Voiced Speech Synthesis.

If $v_l(j) = 1$ and $v_l(j-1) = 1$ and $\frac{|\omega_0(j)-\omega_0(j-1)|}{\omega_0(j)} \geq \rho$ i. e changing pitch,then

$$s_v(n) = s_{v1}(n) + s_{v2}(n) \tag{5.7}$$

where

$$s_{v1}(n) = \sum_{l=1}^{L} w_S(n) M_l(j-1) \cos[\omega_0(j-1)nl + \phi_l(j-1)] \tag{5.8}$$

$$s_{v2}(n) = \sum_{l=1}^{L} w_S(n-N) M_l(j) \cos[\omega_0(j-1)(n-N)l + \phi_l(j)] \tag{5.9}$$

If $v_l(j) = 1$ and $v_l(j-1) = 1$ and $\frac{|\omega_0(j)-\omega_0(j-1)|}{\omega_0(j)} < \rho$ i. e steady pitch,then

$$s_v(n) = \sum_{l=1}^{L} a_l(n) \cos[\theta_l(n)] \tag{5.10}$$

where $a_l(n)$ and $\theta_l(n)$ are the amplitude and phase functions respectively for the l^{th} harmonic,N is the synthesis frame size(=160) and typical $\rho = 0.1$

If the l^{th} harmonic is voiced for current and previous frames and if it is steady state voiced,the energy in this region of the spectrum is reconstructed using interpolation techniques for both amplitudes and phases.To ensure continuity at the beginning and end of a speech frame ,the amplitude function

$a_l(n)$ is linearly interpolated between the estimated values for the current and previous frames using [2, 9]

$$a_l(n) = M_l(j-1) + [M_l(j) - M_l(j-1)]\frac{n}{N} \tag{5.11}$$

Similarly, the phase for the l^{th} harmonic $\theta_l(n)$ can be expressed as

$$\theta_l(n) = \phi_l(j-1) + [l\omega_0(j-1) + \Delta\omega_l(j)]n + [\omega_0(j) - \omega_0(j-1)]\frac{ln^2}{2N} \tag{5.12}$$

where

$$\Delta\omega_l(j) = \frac{1}{N}\left(\Delta\phi_l(j) - 2\pi\lfloor\frac{\Delta\phi_l(j)+\pi}{2\pi}\rfloor\right) \tag{5.13}$$

and

$$\Delta\phi_l(j) = \phi_l(j) - \phi_l(j-1) - [\omega_0(j-1) + \omega_0(j)]\frac{lN}{2} \tag{5.14}$$

The variable $\Delta\omega_l(j)$ is set such that the phase boundary conditions are matched.

5.3.1 Prediction of phases

As seen above, the synthesis of the speech requires information about the phases.Since no phase information is sent from the encoder it has to be predicted at the decoder.The phase information plays a fundamental role in voiced and transition parts of speech segments.To maintain good speech quality ,the phase information must be based on a well defined strategy or model . This can be realised by

1. exploiting phase prediction in voiced speech;

2. maintaining irregular phase structure in unvoiced speech segments.

During steady voicing the speech signal can be thought of as a sequence of periodic impulses which can be decomposed into a set of harmonic sinusoids that add coherently at the time of occurence of each pitch pulse. The phases of these harmonic sinusoids can be predicted using [2]

$$\phi_l(j) = \begin{cases} \psi_l(j) & \text{for } 1 \leq l \leq \frac{L}{4} \\ \psi_l(j) + \frac{L_{uv}\rho_l(j)}{L(j)} & \text{for } \frac{L}{4} < l \leq \max[L(j), L(j-1)] \end{cases}$$

where L_{uv} is the number of unvoiced harmonics,L is the total number of harmonics, ρ_l is a random number generator which are uniformly distributed in the range $[-\pi, \pi]$ and $\psi_l(j)$ is computed as

$$\psi_l(j) = \psi_l(j-1) + [\omega_0(j-1) + \omega_0(j)]\frac{lN}{2} \tag{5.15}$$

This way, all phases corresponding to the voiced harmonics are predicted at the decoder using the previous frame's fundamental frequency and phase and the current frame's fundamental frequency.

5.4 Adaptive post filtering for speech enhancement

The problem of adaptive post filtering for speech enhancement can be thought of as the problem of removing noise (in this case quantisation noise) from a signal.Now Weiner filtering theory gives the optimal linear filter for estimating a signal corrupted with noise.The Weiner filter for estimating a signal corrupted with noise that is uncorrelated with the signal based on minimum mean square error criterion can be written as $\frac{S(\omega)}{S(\omega)+N(\omega)}$ where $S(\omega)$ and $N(\omega)$ are the power spectral densities of the processes $s(t)$ and $n(t)$ respectively.What the equation says is that attenuate those parts of the spectrum where noise power is more so that the overall SNR is improved.This also gives us an idea of how the quantisation should be done.If at the quantiser the (quantisation) noise can be shaped in such a way that the noise is concentrated in a particular region of the spectrum then that noise can be easily removed by the use of a postfilter (at the expense of attenuating the speech component also).

Now we will see how the choice of the particular distortion measure in the previous chapter achieves this objective.Since our distortion measure gives more weight to the LSFs in the formant regions,the noise level in the formant regions will be less but the noise level in the spectral valleys will be more.(Note that this is possible because of the localised spectral sensitivity property of the LSFs).This gives us an idea of how our post filter should look like. It should provide minimum attenuation in the formant regions and maximum attenuation in the spectral valleys.Such a post filter is discussed next.

In the frequency domain post filter we start by weighting the estimated spectral envelope to remove the spectral tilt and produce an even (flatter) spectrum.

$$R_w(e^{j\omega}) = H(e^{j\omega})W(e^{j\omega}) \tag{5.16}$$

Here $H(e^{j\omega})$ is the estimated spectral envelope.

$$H(e^{j\omega}) = \frac{1}{1+\sum_{i=1}^{p} a_k e^{-j\omega k}} \tag{5.17}$$

and

$$W(e^{j\omega)} = \frac{1}{H(e^{\frac{j\omega}{\gamma}})}, 0 \leq \gamma \leq 1 \tag{5.18}$$

and a_k's are the coefficients of the p^{th} order all pole filter and typical $\gamma = 0.5$.

Thus the post filter is given by

$$P_f(e^{j\omega}) = \left(\frac{R_w(e^{j\omega})}{R_{max}}\right)^{\beta}, 0 \leq \beta \leq 1 \tag{5.19}$$

The value of β was taken to be 0.2 and R_{max} is the maximum value of the weighted spectral envelope.

The idea is that at the formant peaks, the normalised spectral envelope will have unity gain and will not be altered by the power β . In the formant nulls , however the fractional values will be pulled up controlling the amount of deemphasis at the formant nulls.

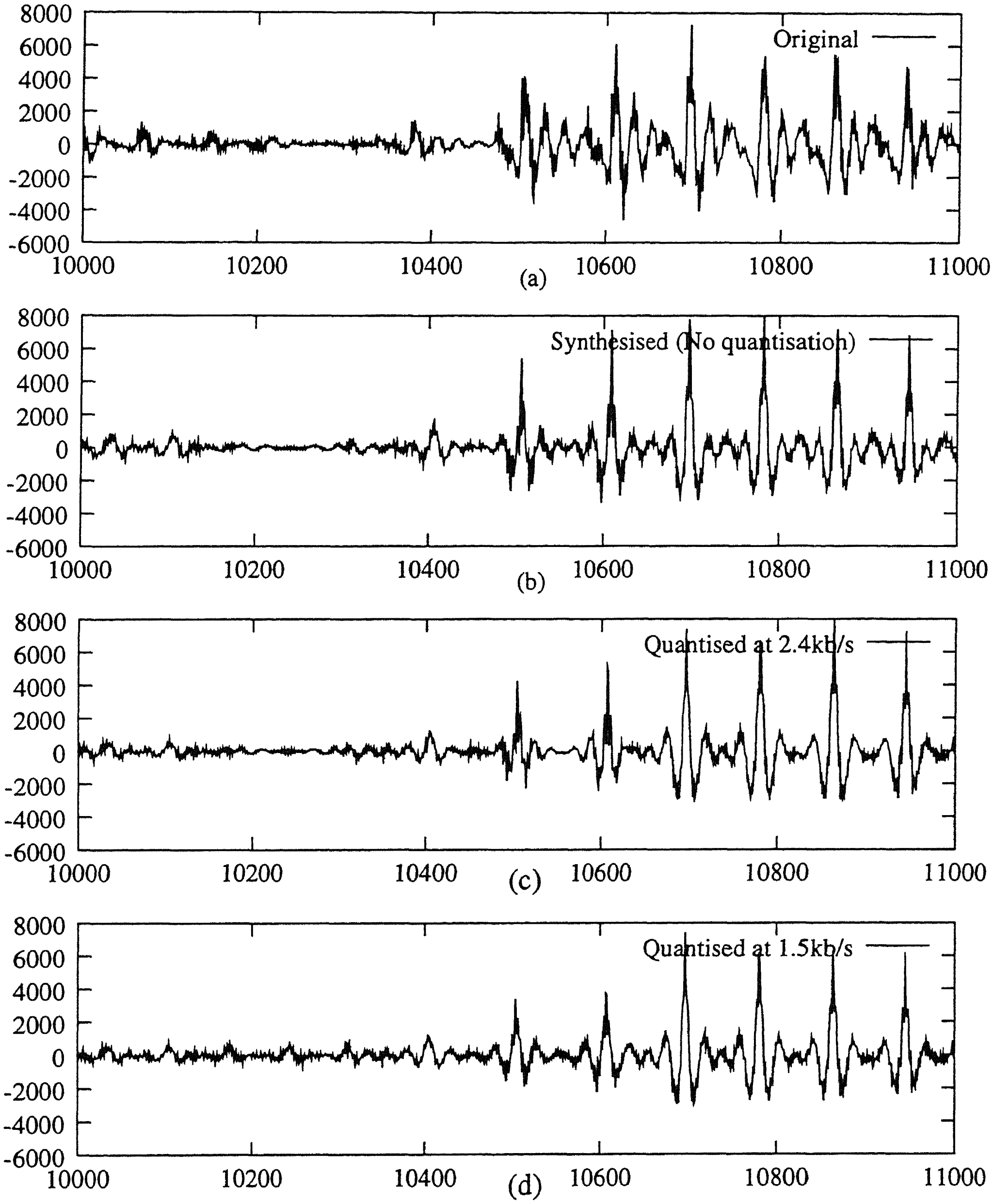

Figure 5.4: (a)Original speech. Synthesied speech (b)No quantisation (c)Bit rate of 2.4kb/s. (d) Bit rate of 1.5kb/s.

Chapter 6

Coder Implementation and Performance

We have studied and implemented the multi band excitation vocoder both at 2.4 kb/s and at 1.5kb/s.The implementation was done in C language on HP-9000 machine.The processor time taken for various steps for 1s of input speech for both 2.4 kb/s & 1.5kb/s is shown in tables 6.1,6.2 respectively.

Since the expensive codebook search is performed only on alternate frames for a 1.5 kb/s coder the processor time required for quantisation in a 1.5kb/s coder is less.

The algorithmic delay of the coder for 2.4kb/s is shown in table 6.3 and for the 1.5kb/s coder is shown in table 6.4. The delay in the analysis is due to the look ahead pitch tracking which should wait for the next two frames to estimate the pitch of the current frame.The delay in synthesis is due to the synthesis window being longer than the frame size.An additional delay n quantisation is encountered in the 1.5kb/s coder due to frame interpolation.

Table 6.1: Processor time for 2.4kb/s coder

Algorithm	*Processor time(s)*
Analysis	30
Quantisation	2.1
Reconstruction	0.14
Synthesis	2.6

The coder was tested on both clean and noisy speech.The synthesised speech was highly intelligible.A certain amount of reverberance was observed even with unquantised parameters of the model. The subjective quality of the speech at 2.4kb/s was not much different from speech synthesised using unquantised parameters but the quality of the speech synthesised at 1.5kb/s was slightly more degraded.

Data from the TIMIT database was used to train the vector quantiser. 25,000 vectors were used to train each of the vector quantiser. Sentences from both inside and outside the training sequence were tested and they did not show much difference in quality showing that the vector quantiser was adequately trained.

Our method of quantising the LSF parameters although efficient was not optimal.Optimal quantisation of LSF parameters can be employed to further improve the performance of the coder.

Table 6.2: Processor time for 1.5kb/s coder

Algorithm	*Processor time(s)*
Analysis	30
Quantisation	1.8
Reconstruction	0.14
Synthesis	2.6

Table 6.3: Algorithmic delay for 2.4kb/s coder

Algorithm	*Delay(ms)*
Analysis	56.25
Quantisation	0.0
Reconstruction	0.0
Synthesis	6.25

Table 6.4: **Algorithmic delay for 1.5kb/s coder**

Algorithm	*Delay(ms)*
Analysis	56.25
Quantisation	20
Reconstruction	20
Synthesis	6.25

Bibliography

[1] D.W.Griffin and J.S.Lim, "Multiband Excitation Vocoder," *IEEE Trans. Acoustics and Speech Signal Processing.*, vol. 36, no. 8, pp. 1223–1245, August 1988.

[2] Digital Voice System Inc,USA, *INMARSAT M Voice codec, Version 3.0*, 1991.

[3] A.M.Kondoz, *Digital Speech Coding for low bit rate communication systems*,Chapter 8, John Wiley., 1991.

[4] K.Paliwal and B.S.Atal, "Efficient quantisation of LPC parameters at 24 bits/frame," *IEEE Trans. Acoustics and Speech Signal Processing.*, vol. 1, no. 1, pp. 3–14, Jan 1993.

[5] N.Sugamura and F.Itakura, "Speech analysis and synthesis methods developed at ECL in NTT-from LPC to LSP," *Speech Communication.*, vol. 5, no. 2, pp. 199–215, June 1986.

[6] Makhoul et.al, "Vector quantisation in speech coding," *Proceedings of the IEEE.*, vol. 73, no. 11, pp. 1551–1588, Nov 1985.

[7] A.Buzo Linde and R.M.Gray, "An algorithm for vector quantiser design," *IEEE Trans On Communications.*, vol. 28, pp. 84–95, Jan 1980.

[8] D.W.Griffin and J.S.Lim, "Signal estimation from modified short time Fourier transform," *IEEE Trans. Acoustics and Speech Signal Processing.*, vol. 32,No 2 ,pp 236-243,April 1984.

[9] R McAulay and T Quatieri, "Speech analysis-synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics and Speech Signal Processing.*, vol. 34, no. 4, Aug 1986.

[10] W.Press et.al, *Numerical Recipes in C*, Cambridge University Press., 1988.

[11] Chen and Gresho, "Adaptive postfiltering for quality improvement," *IEEE Trans. Acoustics and Speech Signal Processing.*, vol. 3, no. 1, Jan 1995.

[12] M.S.Brandstein et.al, "A real time implementation of Improved MBE speech coder," *Proceedings of ICASSP.*,pp 5-8,Apr 1990.

[13] A Gresho, "Advances in speech and audio compression," *Proceedings of the IEEE.*, vol. 82, no. 6, pp. 900–918, Jan 1994.

[14] Schafer and Markel, *Speech Analysis*, IEEE Press., 1976.